

Blockpraktikum  
zur Statistik mit R:  
Lineare Regression

29. September 2011

Till Breuer und Matti Schneider

# Gliederung

- 1 Lineare Zusammenhänge erkennen
  - Korrelationskoeffizienten
  - Korrelation vs. Kausalität
- 2 Lineare Regressionsmodell
  - Methode der kleinsten Quadrate
  - Residuenanalyse und Bestimmtheitsmaß
- 3 Multiple lineare Regression
  - Die Modellierung
  - Der KQ-Schätzer
- 4 Logit Modelle
  - Modellierung binärer Regression
  - Beispiel und Parameterschätzung

# Gliederung

- 1 Lineare Zusammenhänge erkennen
  - Korrelationskoeffizienten
  - Korrelation vs. Kausalität
- 2 Lineare Regressionsmodell
  - Methode der kleinsten Quadrate
  - Residuenanalyse und Bestimmtheitsmaß
- 3 Multiple lineare Regression
  - Die Modellierung
  - Der KQ-Schätzer
- 4 Logit Modelle
  - Modellierung binärer Regression
  - Beispiel und Parameterschätzung

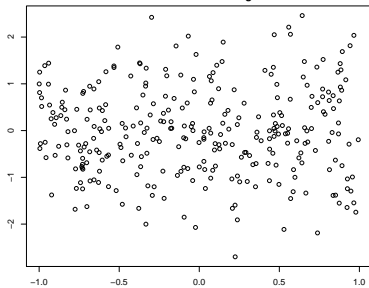
# Lineares Modell

Modellannahme: Zwischen zwei Merkmalen  $X$  und  $Y$  besteht ein *zufällig gestörter* funktionaler Zusammenhang der Form

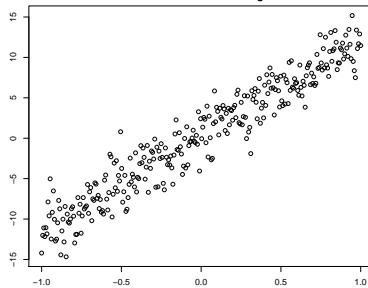
$$Y = f(X) + \epsilon$$

- ▶  $Y$  ist im Wesentlichen eine *deterministische lineare* Funktion von  $X$
- ▶ Der *zufällige Fehler*  $\epsilon$  geht additiv in das Modell ein
- ▶ Gilt  $Y_i = \alpha + \beta X_i + \epsilon_i \rightsquigarrow$  (einfaches) Lineares Regressionsmodell
- ▶ Gilt  $Y_i = \alpha + \beta_1 X_i^1 + \dots + \beta_p X_i^p + \epsilon_i \rightsquigarrow$  Polynomiale Regression
- ▶ Gilt  $Y_i = \alpha + \beta_1 X_{i,1} + \dots + \beta_p X_{i,p} + \epsilon_i \rightsquigarrow$  Multiple lineare Regression
- ▶ Streudiagramm der Daten: Gibt Hinweis auf einen Zusammenhang

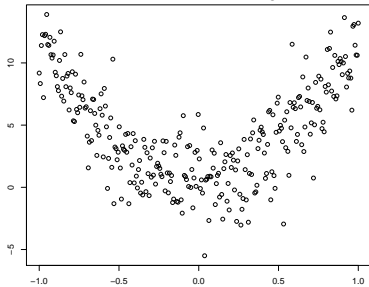
### Kein Zusammenhang



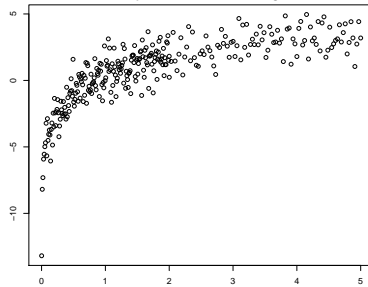
### Linearer Zusammenhang



### Quadratischer Zusammenhang



### Exponentieller Zusammenhang

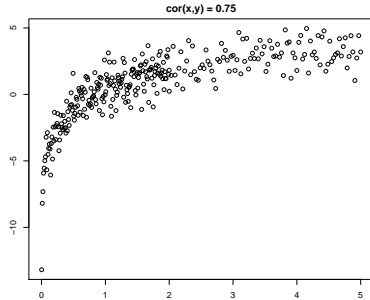
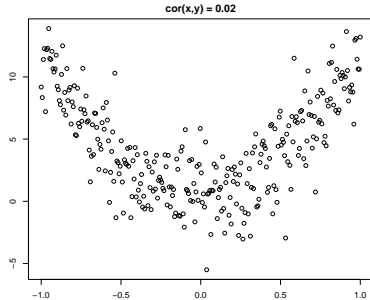
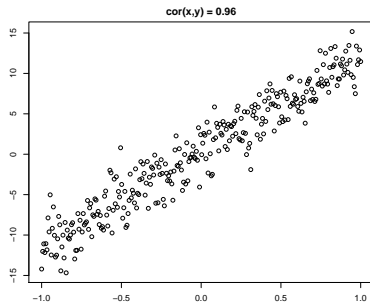
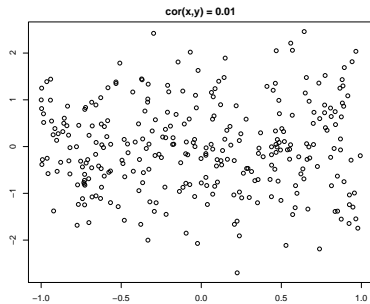


# Bravais-Pearson-Korrelationskoeffizient

Bei linearem Zusammenhang ist der *empirische Korrelationskoeffizient* (auch *Bravais-Pearson-Korrelationskoeffizient*) ein Maß für die Stärke des Zusammenhangs:

$$r := r_{xy} := \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

- ▶  $r > 0$  deutet auf *gleichsinnigen* linearen Zusammenhang (d.h. die Werte liegen um eine Gerade mit positiver Steigung)
- ▶  $r < 0$ : *gegensinniger* linearen Zusammenhang
- ▶  $r = 0$ : unkorreliert, kein linearer Zusammenhang



## Eigenschaften von $r$

- ▶  $r$  ist das empirische Gegenstück zur theoretischen Kovarianz (Ersetzung der (Ko)Varianzen durch empirische Gegenstücke)
- ▶  $r$  nimmt Werte in  $[-1, 1]$  an
- ▶  $|r| = 1$ : Punkte liegen genau auf einer Geraden
- ▶ Berechnung in  $\mathbb{R}$  mit  $\text{cor}(x, y)$

Grobes Einteilungsraster von Korrelationen:

„schwache Korrelation“  $|r| < 0.5$

„mittlere Korrelation“  $0.5 \leq |r| < 0.8$

„starke Korrelation“  $0.8 \leq |r|$



## Spearman'scher Korrelationskoeffizient

Bei Vermutung auf monotonen Zusammenhang  $\leadsto$  bilde *Spearman'schen Korrelationskoeffizient*

$$r_{\text{SP}} = \frac{\sum_{i=1}^n (\text{rg}(x_i) - \overline{\text{rg}}_x)(\text{rg}(y_i) - \overline{\text{rg}}_y)}{\sqrt{\sum_{i=1}^n (\text{rg}(x_i) - \overline{\text{rg}}_x)^2 \sum_{i=1}^n (\text{rg}(y_i) - \overline{\text{rg}}_y)^2}}$$

wobei  $\text{rg}(x_i)$  = Rang von  $x_i$ , und  $\overline{\text{rg}}_x = n^{-1} \sum_{i=1}^n \text{rg}(x_i)$

- ▶  $r_{\text{SP}} > 0$ : gleichsinniger monotoner Zusammenhang  
also: „ $x$  groß  $\Leftrightarrow y$  groß“ sowie: „ $x$  klein  $\Leftrightarrow y$  klein“
- ▶  $r_{\text{SP}} < 0$ : gegensinniger monotoner Zusammenhang
- ▶  $r_{\text{SP}} = 0$ : kein monotoner Zusammenhang
- ▶  $|r_{\text{SP}}| = 1$ : Die Punkte  $(\text{rg}(x_i), \text{rg}(y_i))$ ,  $i = 1, \dots, n$  liegen auf einer Geraden

## Spearman'scher Korrelationskoeff.: Bsp & R-Befehle

Für  $x = (7, 5, 9, 6)$  bzw.  $y = (8, 9, 6, 7)$  gilt

- ▶  $rg(x_1) = 3$  (7 ist drittkleinsten Wert in  $x$ )
- ▶  $rg(x_2) = 1$  (5 ist kleinste Wert in  $x$ )
- ▶ Also:  $rg(x) = (3, 1, 4, 2)$  bzw.  $rg(y) = (3, 4, 1, 2)$
- ▶ Somit:  $r_{SP} = -0.8$

Die zugehörigen R-Befehle lauten

- ▶ `rank(x)` für die Rang-Statistik von  $x$
- ▶ `cor(x, y, method="spearman")` für  $r_{SP}$

# Korrelation vs. Kausalität

- ▶ Korrelation erfasst die Stärke von Zusammenhängen, *keine* Wirkungen oder Wirkungsrichtungen
- ▶ Kausalzusammenhänge lassen sich nur aus inhaltlichen Überlegungen begründen
- ▶ Dies erfordert tiefere Kenntnisse aus dem entspr. Forschungsgebiet
- ▶ Bsp:  $X$  beeinflusst zwar  $Y$ , aber nicht direkt, sondern über ein direktes Merkmal  $Z$
- ▶ Nichtberücksichtigung von  $Z$  führt dann zu falschen Schlüssen

# Scheinkorrelation

Bei 5 Kindern wurde der Wortschatz  $X$  und die Körpergröße  $Y$  gemessen

$x_i$	37	30	20	28	35
$y_i$	130	112	108	114	136

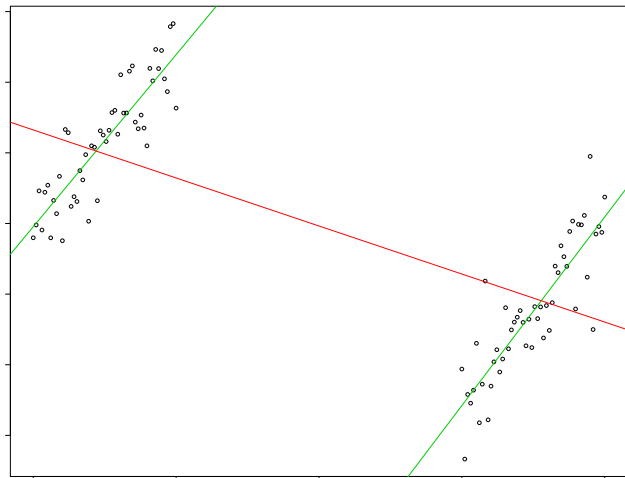
- ▶ Korrelation ist hoch:  $\text{cor}(x,y) = 0.863$
- ▶ Sachlogisch lässt sich eine Beeinflussung nicht begründen
- ▶ Merkmal Alter ( $Z$ ) muss hier berücksichtigt werden

$z_i$	12	7	6	7	13
-------	----	---	---	---	----

- ▶ Es gilt:  $\text{cor}(x,z) = 0.867$  und  $\text{cor}(y,z) = 0.995$

## Verdeckte Korrelation

Nichtbeachtung eines Merkmals kann Korrelation verschleiern oder das Vorzeichen eines Zusammenhanges ändern:



# Gliederung

- 1 Lineare Zusammenhänge erkennen
  - Korrelationskoeffizienten
  - Korrelation vs. Kausalität
- 2 Lineare Regressionsmodell
  - Methode der kleinsten Quadrate
  - Residuenanalyse und Bestimmtheitsmaß
- 3 Multiple lineare Regression
  - Die Modellierung
  - Der KQ-Schätzer
- 4 Logit Modelle
  - Modellierung binärer Regression
  - Beispiel und Parameterschätzung

# (homoskedastische) lineare Regression

Einfachste Modell des linearen Zusammenhangs:

$$y_i = \alpha + \beta x_i + \epsilon_i \quad i = 1, \dots, n$$

wobei

- ▶  $y_1, \dots, y_n$  beobachtbare metrische Zufallsvariablen
- ▶  $x_1, \dots, x_n$  gegebene Realisierungen (*Kontrollparameter*)
- ▶  $\alpha$  und  $\beta$  unbekannte, zu schätzende Parameter sind
- ▶  $\epsilon_1, \dots, \epsilon_n$  (unbeobachtbare) Zufallsvariablen mit:
  - ▶  $\epsilon_1, \dots, \epsilon_n$  sind unabhängig
  - ▶  $\mathbb{E}(\epsilon_i) = 0$  für alle  $i$
  - ▶  $\text{Var}(\epsilon_i) = \sigma^2$  für alle  $i$  (*homoskedastisches Verhalten*)

# Methode der kleinsten Quadrate

- ▶ Ziel: Schätze anhand der Daten  $x$  und  $y$  die Parameter  $\alpha$  und  $\beta$  so, dass die Daten „gut an das Modell angepasst sind“
- ▶ Methode der kleinsten Quadrate: Suche  $\hat{\alpha}$ ,  $\hat{\beta}$  mit

$$\sum_{i=1}^n (x_i - \hat{\alpha} - \hat{\beta}y_i)^2 = \min_{(\alpha, \beta)} \sum_{i=1}^n (x_i - \alpha - \beta y_i)^2$$

- ↪ (summierte) quadratische Abstand zwischen den beobachteten Daten und der Geraden  $y \mapsto \hat{\alpha} + \hat{\beta}y$  ist minimal



# Kleinste-Quadrate-Schätzer

- ▶ Kurvendiskussion der SQT-Funktion  $(\alpha, \beta) \mapsto \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$  liefert die *Kleinste-Quadrate-Schätzer (KQS)*  $\hat{\alpha}$  und  $\hat{\beta}$ :

$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x} \quad \text{und} \quad \hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- ▶ Lineare Regression in R: Aufruf von `lm(y~x)`
- ▶ Wenn man sachlogisch  $\alpha = 0$  annimmt: Aufruf `lm(y~x+0)`
- ▶ `lm()` gibt eine Liste zurück, die man besser immer zwischenspeichert
- ▶ Durch den anschließenden Aufruf von `summary(lm)` erhält man eine detaillierte Ausgabe der Modellanpassung
- ▶ Einfügen der Regressionsgerade in das Streudiagramm: `abline(lm)`

# Residuen

- ▶ Abweichungen  $\widehat{\epsilon}_i$  zwischen Beobachtungen  $y_i$  und den durch das Modell vorhergesagten  $y$ -Werten  $\widehat{y}_i$  nennt man *Residuen*
- ▶ Bei linearer Regression mit KQS:  $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta} \cdot x_i$ , also

$$\widehat{\epsilon}_i = y_i - \widehat{y}_i = y_i - \widehat{\alpha} - \widehat{\beta} \cdot x_i$$

- ▶ Ziel: „Güte des Modells“ anhand der Residuen beurteilen
- ▶ Mögliches Gütemaß: Welcher Anteil der Streuung der  $y_i$  lässt sich durch die Regression erklären? (dieser Wert sollte möglichst groß sein)
- ▶ Gesamtstreuung SQT (*sum of squares total*) ist gegeben durch

$$\text{SQT} = \sum_{i=1}^n (y_i - \bar{y})^2$$

# Streuungszerlegung

- ▶ SQT lässt sich wie folgt zerlegen:

$$\begin{aligned} \text{SQT} &= \text{SQE} + \text{SQR} \\ \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \end{aligned}$$

- ▶ SQE  $\hat{=}$  durch das Modell erklärte Abweichungen
- ▶ SQR  $\hat{=}$  Streuung der Residuen
- ↪ Der Anteil der Streuung der  $y_i$ , die durch die Regression erklärt werden ist durch SQE/SQT gegeben
- ▶ Diesen Wert nennt man das *Bestimmtheitsmaß*  $R^2$

# Bestimmtheitsmaß $R^2$

Für  $R^2$  gilt

$$R^2 = \frac{SQE}{SQT} = 1 - \frac{SQR}{SQT}$$

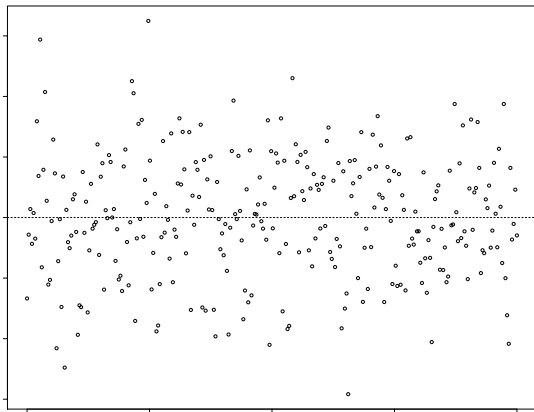
- ▶  $R^2$  liegt zwischen 0 und 1
- ▶ Zusammenhang mit ▶ Bravais-Pearson-Korrelationskoeffizienten  $r$ :

$$R^2 = r^2$$

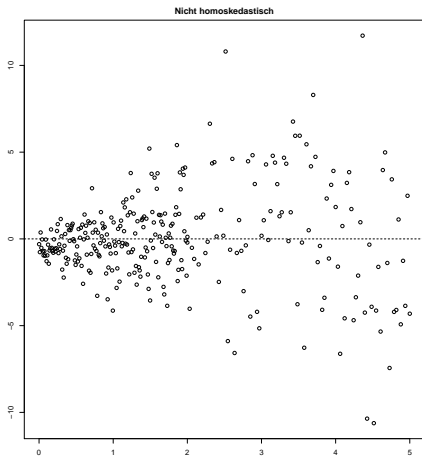
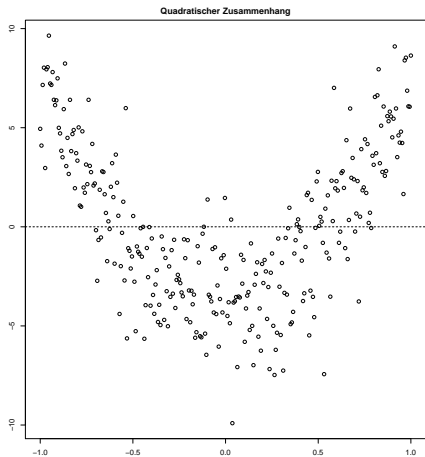
- ▶  $R^2 = 1$  heißt: Daten  $x, y$  liegen auf einer Geraden
- ▶  $R^2 = 0$ : Erklärte Streuung ist 0 (Regressionsmodell ist ungeeignet)

# Residualplots

- ▶ Anhand der Residuen können die Modellannahmen überprüft werden
- ▶ Residuen sollten mit ähnlicher Schwankungsbreite (homoskedastisch) um den Nullpunkt ( $\mathbb{E}(\epsilon_i) = 0$ ) streuen



# Residualplots: Beispiele



## Struktur von `lm()`

- ▶ Betrachte Körpergröße und Alter von 5 Kindern (siehe Folie 2) und erstelle mit `model<-lm(groesse~alter)` ein lineares Modell
- ▶ `model` ist eine Liste, deren Struktur man sich mit `str(model)` anzeigen lassen kann
- ▶ In `model$coefficients` stehen die KQS
- ▶ In `model$residuals` die Residuen
- ▶ Mit `summary(model)` erhält man eine detaillierte Information über das Modell

## Ausgabe von `summary(model)`

```

Call:
lm(formula = y ~ z)

Residuals:
    1      2      3      4      5
-1.2857 -0.4762 -0.7143  1.5238  0.9524

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  86.1429      1.9967   43.14 2.74e-05 ***
z             3.7619       0.2112   17.81 0.000386 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.369 on 3 degrees of freedom
Multiple R-squared:  0.9906, Adjusted R-squared:  0.9875
F-statistic: 317.3 on 1 and 3 DF, p-value: 0.0003857

```

- ▶ KQS:  $\hat{\alpha} = 86.1429$ ,  $\hat{\beta} = 3.7619$  (in der Spalte Estimate)
- ▶  $R^2 = 0.9906$  (Multiple R-squared)
- ▶ Es werden Tests durchgeführt, ob die Parameter  $\alpha$  und  $\beta$  signifikant in das Modell eingehen: Bei vielen Sternchen ist das der Fall, bei keinem Sternchen ist das Modell ungeeignet



Die Aufgaben 8.1. bis 8.7. des Aufgabenblattes  
können jetzt bearbeitet werden.

# Gliederung

- 1 Lineare Zusammenhänge erkennen
  - Korrelationskoeffizienten
  - Korrelation vs. Kausalität
- 2 Lineare Regressionsmodell
  - Methode der kleinsten Quadrate
  - Residuenanalyse und Bestimmtheitsmaß
- 3 Multiple lineare Regression
  - Die Modellierung
  - Der KQ-Schätzer
- 4 Logit Modelle
  - Modellierung binärer Regression
  - Beispiel und Parameterschätzung

# Multipl. Regressionsmodell

Allgemeine Modellannahme:

$$Y_i = \theta_1 \cdot X_{i,1} + \dots + \theta_p \cdot X_{i,p} + \epsilon_i, \quad i = 1, \dots, n$$

bzw. in Matrixnotation

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \underbrace{\begin{pmatrix} X_{1,1} & \dots & X_{1,p} \\ \vdots & \ddots & \vdots \\ X_{n,1} & \dots & X_{n,p} \end{pmatrix}}_{=\mathbf{X}} \cdot \underbrace{\begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix}}_{=\boldsymbol{\theta}} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

mit:

- ▶ *Designmatrix*  $\mathbf{X}$  hat vollen Rang ( $\Rightarrow p \leq n$ )
- ▶ Fehler  $\epsilon_1, \dots, \epsilon_n$  unabh. mit Erwartungswert 0, Varianz  $\sigma^2$

## Spezialfälle des Modells

- ▶ Der Fall  $p = 2$ ,  $X_{i,1} = 1$  für alle  $i$  ist das einfache Lineare Regressionsmodell
- ▶ Den Fall  $X_{i,k} = X_i^k$  nennt man *polynomiales Regressionsmodell*
- ↪ Hier hängt  $Y_i$  polynomial von *einer* Variablen  $X_i$  ab, d.h.

$$Y_i = P(X_i) + \epsilon_i$$

für ein Polynom  $P$  mit Grad  $p$

## KQ-Schätzer für $\theta$

- ▶ Gesucht: „guter“ Schätzer für  $\theta$
- ▶ Auch hier erhält man durch die Methode der kleinsten Quadrate einen KQS  $\hat{\theta} = \hat{\theta}(y)$  für  $\theta$

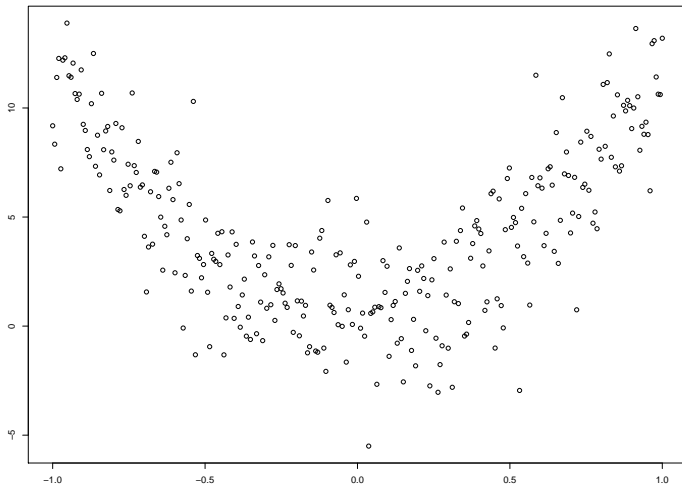
$$\hat{\theta} = (\mathbf{X}^t \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^t \cdot y$$

für eine Beobachtung  $y = (y_1, \dots, y_n)$

- ▶ Berechnung erfolgt numerisch
- ▶ In  $R$  mit `lm(y~x1+x2+...+xp)`

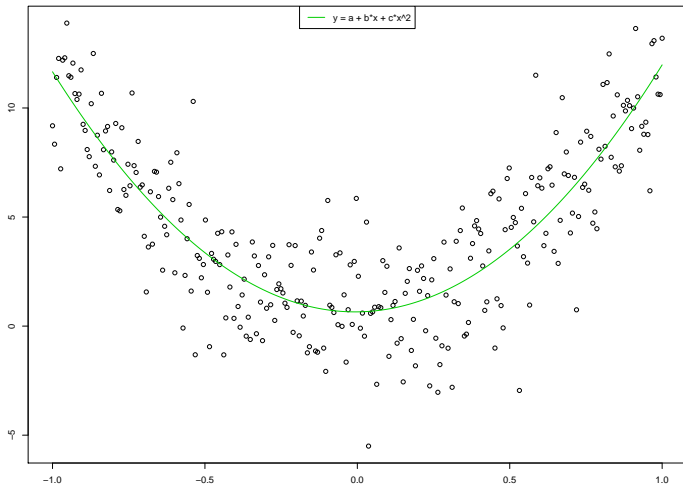
# Beispiel: Polynomielle Regression

Betrachte Beispiel von Folie 5



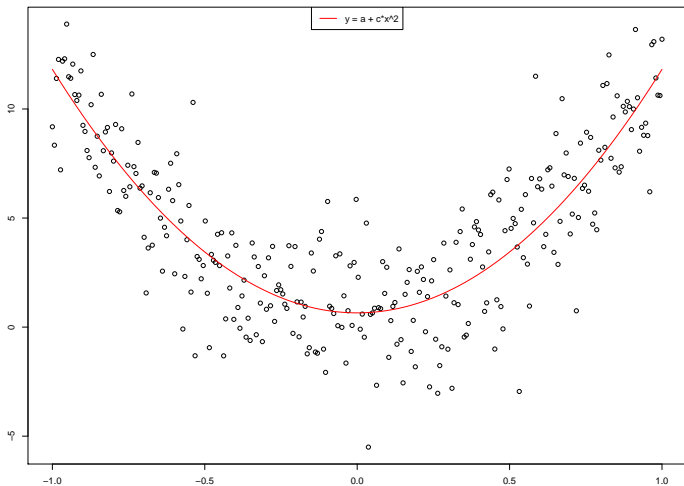
# Beispiel: Polynomielle Regression

Betrachte Beispiel von Folie 5



# Beispiel: Polynomielle Regression

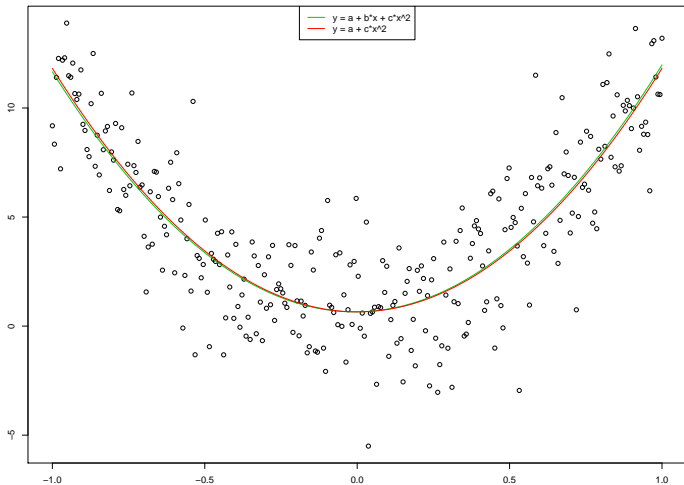
Betrachte Beispiel von Folie 5





# Beispiel: Polynomielle Regression

Betrachte Beispiel von Folie 5



## (cont) Beispiel

- ▶ `summary` Ausgabe für das Modell  $Y_i = a + b \cdot X_i + c \cdot X_i^2 + \epsilon_i$ :

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6527     0.1848   3.532 0.000478 ***
b              0.1586     0.2127   0.746 0.456398
c             11.1651     0.4105  27.200 < 2e-16 ***

```

⇒ Parameter  $b$  geht nicht signifikant in das Modell ein

- ▶ `summary` Ausgabe für das Modell  $Y_i = a + c \cdot X_i^2 + \epsilon_i$ :

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6527     0.1847   3.535 0.000473 ***
c             11.1651     0.4102  27.220 < 2e-16 ***

```

⇒ Modell mit weniger Parametern ist hier vorzuziehen

# Gliederung

- 1 Lineare Zusammenhänge erkennen
  - Korrelationskoeffizienten
  - Korrelation vs. Kausalität
- 2 Lineare Regressionsmodell
  - Methode der kleinsten Quadrate
  - Residuenanalyse und Bestimmtheitsmaß
- 3 Multiple lineare Regression
  - Die Modellierung
  - Der KQ-Schätzer
- 4 **Logit Modelle**
  - **Modellierung binärer Regression**
  - **Beispiel und Parameterschätzung**

# Binäre Regression

- ▶ Lineare Modelle eignen sich gut, wenn die Zielvariable stetig ist (und zumindestens approximativ normalverteilt)
- ▶ In vielen Anwendungen ist Zielvariable nicht stetig, sondern binär
- ▶ Bsp: Antwort auf eine Ja/Nein Frage
- ▶ Ziel einer binären Regression: Modellierung und Schätzung des Effekts der Kovariablen  $X_{i,1}, \dots, X_{i,p}$  auf die (bedingte) Wahrscheinlichkeit

$$\pi_i = \mathbb{P}(Y_i = 1 \mid X_{i,1}, \dots, X_{i,p})$$

# Modellierungsprobleme

Der intuitive Ansatz

$$Y_i = \pi_i + \epsilon_i = \theta_0 + \theta_1 \cdot X_{i,1} + \dots + \theta_p \cdot X_{i,p} + \epsilon_i, \quad i = 1, \dots, n$$

ist ungeeignet:

- ▶  $Y_i$  ist binär, was für die rechte Seite eine zu starke Einschränkung darstellt
- ▶ Fehlervarianz  $\text{Var}(\epsilon_i) = \text{Var}(Y_i | X_{i,1}, \dots, X_{i,p})$  ist nicht homoskedastisch, denn

$$\text{Var}(Y_i | X_{i,1}, \dots, X_{i,p}) = \pi_i \cdot (1 - \pi_i)$$

hängt von  $\theta$  und den Kovariablen  $X_{i,1}, \dots, X_{i,p}$  ab (also nicht unabhängig von  $i$ )

# Modellierung

- ▶ Lösungsansatz: Verknüpfe den *linearen Prädiktor*

$\eta_i = \theta_0 + \theta_1 \cdot X_{i,1} + \dots + \theta_p \cdot X_{i,p}$  mit einer monoton wachsenden Funktion  $h: \mathbb{R} \rightarrow [0, 1]$ , also

$$\pi_i = h(\theta_0 + \theta_1 \cdot X_{i,1} + \dots + \theta_p \cdot X_{i,p})$$

und untersuche das Modell

$$h^{-1}(\pi_i) = \theta_0 + \theta_1 \cdot X_{i,1} + \dots + \theta_p \cdot X_{i,p}$$

- ▶ Als *Responsefunktion*  $h$  wählt man häufig eine Verteilungsfunktion
- ▶ Im Logit-Modell:  $h(t) = \frac{\exp(t)}{1+\exp(t)}$  (logistische Verteilungsfunktion)
- ▶ Umkehrfunktion  $h^{-1}(t) = \log\left(\frac{t}{1-t}\right)$  heißt auch *Logit-Funktion*

## Modelle für die Chancen

- ▶ Für die *Chancen (odds)*

$$\frac{\pi_i}{1 - \pi_i} = \frac{\mathbb{P}(Y_i = 1 \mid X_{i,1}, \dots, X_{i,p})}{\mathbb{P}(Y_i = 0 \mid X_{i,1}, \dots, X_{i,p})}$$

erhält man (im Logit Modell!) das multiplikative Modell

$$\frac{\mathbb{P}(Y_i = 1 \mid X_{i,1}, \dots, X_{i,p})}{\mathbb{P}(Y_i = 0 \mid X_{i,1}, \dots, X_{i,p})} = \exp(\theta_0) \cdot \exp(\theta_1 X_{i,1}) \cdot \dots \cdot \exp(\theta_p X_{i,p})$$

- ⇒ Für die logarithmierten Chancen (*log-odds*) erhält man ein lineares Modell

## Interpretation

Bei Erhöhung der Kovariablen  $X_{i,k}$  um 1 ändert sich das Verhältnis der Chancen um  $\exp(\theta_k)$

$$\frac{\mathbb{P}(Y_i = 1 \mid X_{i,k}, \dots)}{\mathbb{P}(Y_i = 0 \mid X_{i,1}, \dots)} = \exp(\theta_k) \cdot \frac{\mathbb{P}(Y_i = 1 \mid X_{i,1} + 1, \dots)}{\mathbb{P}(Y_i = 0 \mid X_{i,1} + 1, \dots)}$$

Also:

- ▶  $\theta_k > 0$ : Chance wird größer
- ▶  $\theta_k < 0$ : Chance wird kleiner
- ▶  $\theta_k = 0$ : Chance bleibt gleich



## Beispiel: Tod durch Herzversagen

- ▶ Zielvariable: Patient stirbt an Herzversagen (0/1 codiert)
  - ▶ Kovariablen: Alter, Geschlecht, Cholesterinspiegel, ...
  - ▶ Ist  $\theta_1 = 0.08$ : Chance an Herzversagen zu sterben erhöht sich bei einem 10 Jahre älterem Patienten um den Faktor  $\exp(10 \cdot 0.08) \approx 2.2$
  - ▶ Vorsicht: Bei unterschiedlichen Wahlen von  $h$  können die geschätzten Werte für  $\theta$  stark abweichen, die Verhältnisse  $\theta_2/\theta_1$  usw. bleiben jedoch nahezu konstant
- ⇒ Es kann festgestellt werden, welche Kovariable den größten Einfluss hat

# Schätzung der Parameter

- ▶  $\theta$  wird *nicht* durch die Methode der kleinsten Quadrate geschätzt
- ▶ Mit einem Maximum-Likelihood-Ansatz ( $Y_i$  ist ja binär) lässt sich (numerisch) ein guter Schätzer für  $\theta$  bestimmen
- ▶ In R erhält man diese mit  
`glm(Y ~ x1 + ... + xp, family = binomial(link="logit"))`